Statistica Sinica Preprint No: SS-2022-0198	
Title	Approximating Optimal SMC Proposal Distributions in Individual-Based Epidemic Models
Manuscript ID	SS-2022-0198
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0198

Complete List of Authors	Lorenzo Rimella, Christopher Jewell and Paul Fearnhead
Corresponding Authors	Lorenzo Rimella
E-mails	l.rimella@lancaster.ac.uk

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Approximating optimal SMC proposal distributions in individual-based epidemic models

Lorenzo Rimella, Christopher Jewell and Paul Fearnhead

Lancaster University

Abstract: Many epidemic models are naturally defined as individual-based models, in which we track the state of each individual within a susceptible population. However, inference for individual-based models is challenging because of the high-dimensional state-space of such models, which increases exponentially with the population size. Here, we consider sequential Monte Carlo algorithms for inference for individual-based epidemic models, where we make direct observations of the state of a sample of individuals. Standard implementations, such as the bootstrap filter and auxiliary particle filter, are inefficient, owing to a mismatch between the proposal distribution of the state and future observations. We develop new efficient proposal distributions that consider future observations, leveraging the following properties: (i) we can analytically calculate the optimal proposal distribution for a single individual, given future observations and the future infection rate of that individual; and (ii) the dynamics of individuals are independent if we condition on their infection rates. Thus, we construct estimates of the future infection rate for each individual, and then use an independent proposal for the state of each individual, given this estimate. Empirical results show

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

orders of magnitude improvement in efficiency of the sequential Monte Carlo sampler for both SIS and SEIR models.

Key words and phrases: individual-based model; proposal distribution sequential Monte Carlo.

1. Introduction

Dynamical disease transmission models are increasingly being used to inform disease control policy related to both human and livestock outbreaks, for example, the SARS and H1N1 pandemic influenza in humans (Zhou et al., 2004), avian influenza in poultry (Van der Goot et al., 2005), and foot-and-mouth disease in cloven-hoofed livestock (Zhou et al., 2004; Jewell et al., 2009). Most recently, such models have been central in informing national-level decisions on social distancing and vaccination strategies for the SARS-CoV-2 pandemic (Brooks-Pollock et al., 2021; Funk et al., 2020). In addition to managing outbreaks, such models are useful for studying the dynamics of endemic diseases, with the ability to explain random fluctuations around an otherwise stable case incidence, particularly in highly heterogeneous populations (Britton, 2010).

In essence, disease transmission models belong to the class of state-transition models described by a directed (though not necessarily acyclic)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

graph. For example, the susceptible-infected-susceptible (SIS) model proposes individuals as existing as "susceptible" or "infected", and individuals are allowed to transition from either state to the other. In a stochastic setting, it is natural to assume that an individual in a population experiences a hazard rate of progressing from some source state to a destination state. This setup has particular relevance when the transition hazard rates depend on both individuals' characteristics as well as the characteristics of their relationship with each other. Many applications demand individual-level granularity, particularly when observations are of specific individuals or when disease interventions are targeted to particular individuals (Chapman et al., 2020; Jewell et al., 2009; Cocker et al., 2022).

However, inference for such models is challenging, owing to the partial or total censoring of transition events, for which the state-space increases exponentially with the population size. For example, in an SIS model, we may have noisy observations in which individuals exist in either the S or I states at particular times, but no direct observation of when state transitions occur.

Following Rimella et al. (2022), we consider sequential Monte Carlo (SMC) methods for inference for such models. We show that standard SMC implementations (Gordon et al., 1993; Pitt and Shephard, 1999) are

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

inefficient for individual-based epidemic models. In particular, they struggle to propose states for all individuals that are consistent with future observations. Ju et al. (2021) consider how to improve the efficiency of the SMC for individual-based epidemic models. However, they consider observations of, for example, the number of infected individuals, and their approach does not obviously apply to the observation models we consider here.

To improve the efficiency of the SMC, we develop a novel proposal distribution that takes into account future observations. The key idea is based on two properties of the dynamics of individual-based epidemic models. First, it is tractable to calculate the conditional distribution of the state of a single individual, given future observations and the future infection rate of the individual. This can be calculated using standard recursions for finite-state hidden Markov models (Rabiner and Juang, 1986) together with the fact that the state-space for a single individual is small (e.g., two for an SIS model or four for a susceptible-exposed-infected-removed (SEIR) model). Second, there is a form of conditional independence across individuals: if we condition on the future infection rates for each individual, then the dynamics of the state for individuals are independent of each other. In the models, we consider that the infection rate for each individual depends only on the total number of infectious individuals. Thus, we can use

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

the ideas of Whiteley and Rimella (2021) to estimate the future number of infectious individuals. Then, conditioning on this estimate and the corresponding infection rates for each individual, we have a proposal distribution that is independent across individuals. Furthermore, for each individual, the proposal distribution is equal to the true conditional distribution of the state, given the estimated future infection rates and observations for that individual.

The computational cost of using this proposal is proportional to the number of time steps at which we have future observations. In practice, we can implement the proposal distribution by conditioning only on future observations over a suitable time window. We show empirically that using such a distribution can lead to an order of magnitude improvement in Monte Carlo efficiency, even after accounting for the increased computational cost.

2. Preliminaries

We use bold lowercase letters for vectors, for example, $\mathbf{a}$, and bold uppercase letters for matrices, for example, $\mathbf{A}$. We use $\mathbf{A}^{(i,j)}$ for the (i,j)th element of $\mathbf{A}$ and we use $\mathbf{A}^{(i,\bullet)}$ (or $\mathbf{A}^{(\bullet,j)}$) to represent the column vector given by the (i)th column (or the (j)th row) of matrix $\mathbf{A}$. We use $\circ$ and $/$ to denote the elementwise product and ratio, respectively, between vectors or matrices,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

and $\mathbf{1}_M$ denotes the M-dimensional vector of ones. Given $t, s \in \mathbb{N}$ with $t > s$, we use $[s : t]$ for the set $\{s, \ldots, t\}$, for example, for $t \in \mathbb{N}$, we use $y_{[1:t]}$ for $\{y_1, \ldots, y_t\}$. The notation for the main probability distributions is reported in Table 1.

Distribution	Categorical	Bernoulli	Binomial	Gaussian	Uniform	Multinomial
Notation	$\mathcal{C}at_M(i\|\mathbf{p})$	$\mathcal{B}e(i\|q)$	$\mathcal{B}in(i\|N, q)$	$\mathcal{N}(a\|\mu, \sigma^2)$	$\mathcal{U}nif(q\|a, b)$	$\mathcal{M}ult(\mathbf{c}\|N, \mathbf{p})$

Table 1: Notation table for probability mass and density functions.

3. Model

3.1 Individual-based epidemic models

We consider individual-based models defined as follows: the number of compartments M, the population size N, the initial probability of an individual being assigned to a compartment $(\mathbf{p}_{n,0})_{n\in[1:N]}$, and the probability of an individual transitioning from one compartment to another $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$, where the stochastic transition matrix $\mathbf{K}_{n,\bullet}$ is defined as a function of an M-dimensional vector $\mathbf{c}$, that is, $\mathbf{c} \to \mathbf{K}_{n,\mathbf{c}}$. In practice, $\mathbf{c}^{(i)}$ is the number of individuals in compartment i, and so the transition matrix $\mathbf{K}_{n,\mathbf{c}}$ depends only on the state of the compartment. However, more general versions are possible, and are briefly discussed in Section 6 (e.g., spatial models). We

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

use $(\mathbf{x}_t)_{t\geq 0}$ for the population state, and $(\mathbf{c}_t)_{t\geq 0}$ for the compartments state, as follows:

Time 0: $\mathbf{x}_0^{(n)} \sim \mathcal{C}at_M(\bullet|\mathbf{p}_{n,0})$, for $n \in [1:N]$, and

$\mathbf{c}_0^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_0^{(n)}}(i)$, for $i \in [1:M]$;

Time t: $\mathbf{x}_t^{(n)}|\mathbf{x}_{t-1} \sim \mathcal{C}at_M\left(\bullet\middle|\mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\bullet)}\right)$, for $n \in [1:N]$, and

$\mathbf{c}_t^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_t^{(n)}}(i)$, for $i \in [1:M]$.

SIS example We can make the SIS model heterogeneous by following the construction in Ju et al. (2021). Suppose that we have $d \in \mathbb{N}$ covariates for each individual. We can then define $(\mathbf{w}_n)_{n\in[1:N]}$ as the collection of d-dimensional vectors of individual-specific covariates, from which we can compute, for $n \in [1:N]$,

$$\mathbf{p}_{n,0} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \\ \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \end{bmatrix}, \quad \mathbf{K}_{n,c} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(2)}}{N} & \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(2)}}{N} \\ \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} & 1 - \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} \end{bmatrix}$$

with $\beta_0 \in \mathbb{R}^d$ and $\beta_\lambda, \beta_\gamma \in \mathbb{R}^d$. In this model, we have individual-specific probabilities of infection and recovery.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

3.2 Observation model

The observation process is denoted by $(\mathbf{y}_t)_{t\geq 1}$ and, given $(\mathbf{q}_{n,t})_{n\in[1:N],t\geq 1}$, with $\mathbf{q}_{n,t} \in [0,1]^M$, we generate observations per each time step t as follows:

$$\mathbf{y}_t^{(n)} = \mathbf{x}_t^{(n)}\mathbf{r}_t^{(n)} \text{ with } \mathbf{r}_t^{(n)} \sim \mathcal{B}e\left(\bullet \middle| \mathbf{q}_{n,t}^{(\mathbf{x}_t^{(n)})}\right) \text{ for } n \in [1:N], \tag{3.1}$$

which we refer to as the "granular observations model". Note that $\mathbf{y}_t^{(n)} \in [0:M]$, meaning that we either report the state of individual n as it is ($\mathbf{y}_t^{(n)} = \mathbf{x}_t^{(n)}$), or we do not report it at all ($\mathbf{y}_t^{(n)} = 0$). This model includes observations from random samples of the population, where each component of $\mathbf{q}_{n,t}$ is the same and equal to the probability that individual n is included in the sample at time t, as well as situations in which observations are made preferentially for certain states (e.g., observations of infected farms for foot-and-mouth disease). To simplify the notation and derivations, we focus on individual homogeneous reporting rates, that is, $\mathbf{q}_{n,t} = \mathbf{q}_t$, and under this assumption, $\sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_t^{(n)}}(i) \sim \mathcal{B}in(\bullet|\mathbf{c}_t^{(i)}, \mathbf{q}_t^{(i)})$, for any $i \in [1:M]$, which recovers the binomial observation model (Whiteley and Rimella, 2021; Ju et al., 2021).

SIS example Given $\mathbf{q}_t \in [0,1]^2$, $\mathbf{q}_t^{(1)}$ is the probability of reporting a susceptible, and $\mathbf{q}_t^{(2)}$ is the probability of reporting an infected.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

3.3 Inference in individual-based models with granular observations

In epidemiology, we are interested in inferring both the unknown state of the population $\mathbf{x}_t$ and the parameters of the epidemic θ. Given the time horizon t, the individual-based model with granular observation $(\mathbf{x}_s, \mathbf{y}_s)_{s\in[1:t]}$ is, by construction, a hidden Markov model (HMM). We can compute the filtering distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)$ and marginal likelihood $p(\mathbf{y}_{[1:s]}|\theta)$ using the forward algorithm (Rabiner and Juang, 1986). Then, we can infer the parameters using, for example, the EM algorithm (Yang et al., 2017).

The forward algorithm requires marginalizing over the whole state-space, making it unfeasible for our individual-based model, where marginalizations are $\mathcal{O}(M^N)$. As an alternative, we can use sequential Monte Carlo (SMC) algorithms to obtain particle approximations of $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)$ and $p(\mathbf{y}_{[1:s]}|\theta)$ (Ionides et al., 2006; Kucharski et al., 2020) at a cost that is linear in the number of particles and the time horizon. Given a number of particles $P \in \mathbb{N}$, at each time step s, an SMC algorithm proposes instances $(\mathbf{x}_s^p)_{p\in[1:P]}$ of the latent process $(\mathbf{x}_s)_{s\in[1:t]}$ based on the proposal distribution $q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]})$, with $q(\mathbf{x}_0|\mathbf{x}_{-1}, \mathbf{y}_{[1:t]}) \coloneqq q(\mathbf{x}_0|\mathbf{y}_{[1:t]})$, and $q(\mathbf{x}_0|\mathbf{y}_{[1:t]})$, proposal distribution at time $s = 0$, and assigns weights $(w_s^p)_{p\in[1:P]}$ to the particles to produce an importance sample that approximates the filtering

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

distribution. Before moving to the next step, the algorithm uses a resampling scheme $r_s(i)$, a distribution over the indeces of the particles $[1:P]$, to discard low-weight particles. Finally, the procedure generates particle estimates of the filtering distribution, $p(\mathbf{x}_s|\mathbf{y}_{[1:s]},\theta) \approx (\sum_{\tilde{p}\in[1:P]} w_s^p)^{-1} \sum_{p\in[1:P]} w_s^p \delta_{\mathbf{x}_s^p}(\mathbf{x}_s)$, and the marginal likelihood, $p(\mathbf{y}_{[1:t]}|\theta) \approx \prod_{s\in[1:t]} \sum_{p\in[1:P]} w_s^p$.

The performance of SMC algorithms depends heavily on the proposal distribution $(q(\mathbf{x}_s|\mathbf{x}_{s-1},\mathbf{y}_{[1:t]}))_{s\in[0:t]}$ and the resampling scheme $(r_s(i))_{s\in[0:t]}$, where poor choices might lead to high variance of the marginal likelihood estimator, particles/weights degeneracy, and even observation mismatches, which might cause the algorithm to fail. The bootstrap particle filter (BPF) (Gordon et al., 1993; Candy, 2007) proposes new particles based on the transition kernel, and resamples according to the current weights; that is, $q(\mathbf{x}_s|\mathbf{x}_{s-1},\mathbf{y}_{[1:t]}) = p(\mathbf{x}_s|\mathbf{x}_{s-1},\theta)$, with $q(\mathbf{x}_0|\mathbf{y}_{[1:t]}) = p(\mathbf{x}_0|\theta)$, and $r_s(i) = \mathcal{C}at_P\left(i|\left[w_s^1,\ldots,w_s^P\right]\right)$. The BPF is known to perform poorly in high-dimensional scenarios (Bickel et al., 2008) and with informative observations, especially when the simulated particles have to match certain paths. An easy fix is to include the information from the current observations in the proposal distributions, thus avoiding a mismatch at the current time step when proposing new particles. The resulting algorithm is called the auxiliary particle filter (APF) (Pitt and Shephard, 1999; Carpenter et al., 1999;

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Johansen and Doucet, 2008), where $q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}) = p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_s, \theta)$, with $q(\mathbf{x}_0|\mathbf{y}_{[1:t]}) = p(\mathbf{x}_0|\theta)$, and $r_s(i) = \mathcal{C}at_P\left(i|\left[w_s^1, \ldots, w_s^P\right]\right)$.

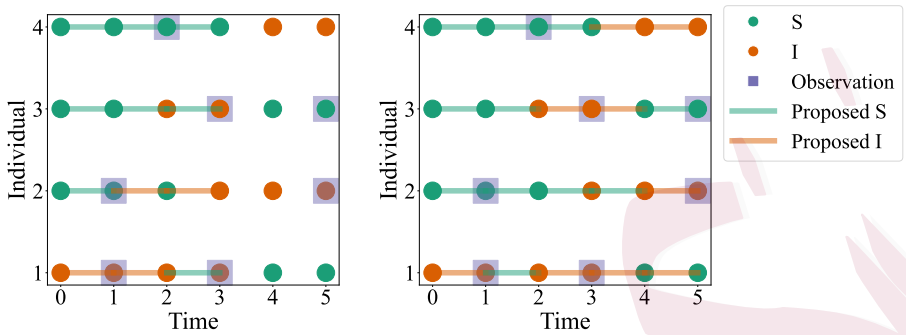

Figure 1: Illustration of the BPF (left) and the APF (right) in an SIS scenario. Colored dots show the state of each individual, with green for susceptible and red for infected. Dots in gray squares are observations. Horizontal lines from $s-1$ to s are used for the proposed states in s.

In Figure 1, we graphically compare between BPF and APF in an individual-based model. The BPF fails after three iterations, because the proposed particle does not match the observed state for individuals 1 and 3; we observe $\mathbf{y}_3^{(1)} = 2$ and $\mathbf{y}_3^{(3)} = 2$, but the BPF proposes $(\mathbf{x}_3^p)^{(1)} = 1$ and $(\mathbf{x}_3^p)^{(3)} = 1$ (green lines). In contrast, the APF proposes particles that are constrained to match the observation, because it includes the current data in the proposal. However, the APF's proposal is still inefficient, because it does not take into account future observations. In Figure 1, this is seen by

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

it tending to propose a switch to the infected state immediately before the observation of an infected individual, whereas an individual often becomes infected one or more time steps earlier. For more complicated models, such as the SEIR model we consider in Section 5.2, the APF can also suffer from mismatch, because the transition to an observed state may not be possible for the current state of a particle.

In the next section, we show how to build, for any $s \in [0:t]$, an approximation of $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ for an individual-based model with granular observations. Given that computing $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ requires $(\mathbf{c}_{\tilde{s}})_{\tilde{s}\in[s:t-1]}$, the main idea is to approximate $(\mathbf{c}_{\tilde{s}})_{\tilde{s}\in[s:t-1]}$ using the expectation of a pre-computed multinomial distribution (Whiteley and Rimella, 2021), and then to propagate backward the observation $\mathbf{y}_{[s:t]}$ to inform the proposal in s.

4. Optimal proposal distributions for individual-based models

The optimal proposal for an SMC is $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$, which we can compute recursively as follows:

Time t: $p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) = \sum_{\mathbf{x}_t\in[1:M]^N} p(\mathbf{y}_t|\mathbf{x}_t, \theta)p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta)$ and

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta) = \frac{p(\mathbf{y}_t|\mathbf{x}_t,\theta)p(\mathbf{x}_t|\mathbf{x}_{t-1},\theta)}{p(\mathbf{y}_t|\mathbf{x}_{t-1},\theta)};$$

Time s: $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1}, \theta) = \sum_{\mathbf{x}_s\in[1:M]^N} p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)p(\mathbf{y}_s|\mathbf{x}_s, \theta)p(\mathbf{x}_s|\mathbf{x}_{s-1}, \theta)$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

$$p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta) = \frac{p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s,\theta)p(\mathbf{y}_s|\mathbf{x}_s,\theta)p(\mathbf{x}_s|\mathbf{x}_{s-1},\theta)}{p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)};$$

Time 0: $p(\mathbf{y}_{[1:t]}, \theta) = \sum_{\mathbf{x}_0 \in [1:M]^N} p(\mathbf{y}_{[1:t]}|\mathbf{x}_0, \theta)p(\mathbf{x}_0, \theta)$ and

$$p(\mathbf{x}_0|\mathbf{y}_{[1:t]}, \theta) = \frac{p(\mathbf{y}_{[1:t]}|\mathbf{x}_0,\theta)p(\mathbf{x}_0,\theta)}{p(\mathbf{y}_{[1:t]}|\theta)}.$$

See Fearnhead (2008) for a review of the optimal proposal for importance sampling, Chopin et al. (2020) for a discussion on the optimal proposal distribution for particle filters and Whiteley and Lee (2014) for a more technical discussion.

A marginalization over the whole state-space is required, resulting in a computational cost of $\mathcal{O}(M^N)$ per step. Observe that at the beginning of the recursion, we can exploit the factorization over the individuals at time t, that is, the components of $\mathbf{x}_t$, of the transition kernel and the emission distribution to reduce the computational cost of the marginalization to $\mathcal{O}(NM)$:

$$p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) = \prod_{n \in [1:N]} \sum_{\mathbf{x}_t^{(n)} \in [1:M]} \mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)}, \mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1 - \mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)}. \tag{4.1}$$

Note that $\mathbf{y}_t^{(n)}$ is not conditionally independent given $\mathbf{x}_{t-1}^{(n)}$, because of the dependence of the transition kernel on the compartments state $\mathbf{c}_{t-1}$. This breaks the computational trick, because we cannot express $p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)$ as a product over the individuals at time $t-1$, and so the cheap marginalization

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

has to be repeated for each state of $\mathbf{x}_{t-1}$, leading to $\mathcal{O}(NM^{N+1})$. However, $\mathbf{y}_t^{(n)}$ is conditionally independent given $\mathbf{x}_{t-1}^{(n)}$ and $\mathbf{c}_{t-1}$, meaning that if an estimate of $\mathbf{c}_{t-1}$ is available a priori, the factorization is preserved, and the same trick can be iterated in the subsequent time steps.

4.1 A priori estimates of the compartments state

Whiteley and Rimella (2021) propose an efficient way of approximating the smoothing distribution $p(\mathbf{c}_s|\mathbf{y}_{[1:t]}, \theta)$ using a multinomial distribution $\mathcal{M}ult(\mathbf{c}_s|N, \mathbf{m}_{s|t})$, with the parameters computed recursively using a forward and a backward step through the data at a computational cost $\mathcal{O}(tM^3)$. In the multinomial approximation, there are two key assumptions: the homogeneity of the individuals, and a binomial observation model of the form $\mathcal{B}in(\bullet|\mathbf{c}_s^{(i)}, \mathbf{q}_s^{(i)})$. We can recover homogeneity in the individual-based model with granular observations by defining the mean initial distribution $\bar{\mathbf{p}}_{n,0}$ and the mean transition kernel $\bar{\mathbf{K}}_{\mathbf{c}_s}$:

$$\bar{\mathbf{p}}_{n,0}^{(i)} := \frac{1}{N} \sum_{n\in[1:N]} \mathbf{p}_{n,0}^{(i)}, \text{ for } i \in [1:M],$$

$$\bar{\mathbf{K}}_{\mathbf{c}_s}^{(i,j)} := \frac{1}{N} \sum_{n\in[1:N]} \mathbf{K}_{n,\mathbf{c}_s}^{(i,j)}, \text{ for } i,j \in [1:M].$$

Note that recovering homogeneity by approximating the individuals' transition kernel using an average is also a key step in Ju et al. (2021), where the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

transition probabilities are approximated by averaging over the individuals to avoid an exponential computational cost in the population size.

We already have $\sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_s^{(n)}}(i) \sim \mathcal{B}in(\bullet|\mathbf{c}_s^{(i)}, \mathbf{q}_s^{(i)})$, for $i \in [1:M]$, from which we can define the cumulative observations per compartment as the vector $\mathbf{o}_s$, with components $\mathbf{o}_s^{(i)} := \sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_s^{(n)}}(i)$.

Using the aforementioned approximate dynamic and observation model, the multinomial approximation in Whiteley and Rimella (2021) scans the data forward and backward, and computes multinomial approximations of the filtering and smoothing distributions (the full algorithm is reported in the Supplementary Material). The forward pass consists of a prediction step and an update step preserving the multinomial form, precisely, starting from $\mathbf{m}_{0|0} := \bar{\mathbf{p}}_{n,0}$, we have

$$\mathbf{m}_{s-1|s} := \left(\mathbf{m}_{s-1|s-1}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_{s-1}}\right)^{\mathrm{T}}, \quad \mathbf{m}_{s|s} := \frac{\mathbf{o}_s}{N} + \left(1 - \frac{\mathbf{1}_M^{\mathrm{T}} \mathbf{o}_s}{N}\right) \frac{\mathbf{m}_{s-1|s} \circ (\mathbf{1}_M - \mathbf{q}_s)}{1 - \mathbf{m}_{s-1|s}^{\mathrm{T}} \mathbf{q}_s},$$

which yields an approximation for the filtering distribution $p(\mathbf{c}_s|\mathbf{y}_{[1:s]}) \approx \mathcal{M}ulti(\mathbf{c}_s|N, \mathbf{m}_{s|s})$. The backward pass implements the following reverse kernel and applies it backward:

$$\mathbf{L}_s := \left\{ \left[(\mathbf{m}_{s|t} \mathbf{1}_M^{\mathrm{T}}) \circ \bar{\mathbf{K}}_{\mathbf{m}_s} \right] / \left[\mathbf{1}_M (\mathbf{m}_{s|t}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_s}) \right] \right\}^{\mathrm{T}}, \quad \mathbf{m}_{s|t} := \left(\mathbf{m}_{s+1|T}^{\mathrm{T}} \mathbf{L}_s\right)^{\mathrm{T}},$$

outputting the M-dimensional probability vector $\mathbf{m}_{s|t}$, and so approximating the smoothing distribution with $p(\mathbf{c}_s|\mathbf{y}_{[1:t]}, \theta) \approx \mathcal{M}ulti(\mathbf{c}_s|N, \mathbf{m}_{s|t})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Given the multinomial approximations, we can approximate the compartments state with

$$\mathbf{c}_s \approx \mathbb{E}_{\mathcal{M}ult(\mathbf{c}_s|N,\mathbf{m}_{s|t})}(\mathbf{c}_s) = N\mathbf{m}_{s|t}.$$

We have imposed a restriction on the emission distribution by assuming a uniform reporting probability for all individuals. However, our approach can be extended to accommodate a more general scenario in which $\mathbf{q}_{t,n}$ varies with n. To do so, we can compute the mean reporting rate $\bar{\mathbf{q}}_t := (N)^{-1}\sum_{n\in[1:N]}\mathbf{q}_{t,n}$ when running Whiteley and Rimella (2021), and then substitute back $\mathbf{q}_{t,n}$ when computing the approximation to the optimal proposal.

4.2 Approximate optimal proposals for individual-based models

Conditioning on $\mathbf{c}_{\tilde{s}} = N\mathbf{m}_{\tilde{s}|t}$, for $\tilde{s} \in [s:t]$, makes the individuals evolve independently from each other, and so it allows an analytical computation of $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)$ at a cost $\mathcal{O}(NM)$. Starting again from (4.1):

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_{t-1},\theta) &\approx \prod_{n\in[1:N]} \sum_{\mathbf{x}_t^{(n)}\in[1:M]} \mathbf{K}_{n,N\mathbf{m}_{t-1|t}}^{(\mathbf{x}_{t-1}^{(n)},\mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1-\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)} \\ &=: \prod_{n\in[1:N]} \boldsymbol{\xi}_{n,t-1}^{(\mathbf{x}_{t-1}^{(n)})}, \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

where we define the quantities $\boldsymbol{\xi}_{n,t-1}$ for each individual n as the approximate probability of observing the future observation $\mathbf{y}_t^{(n)}$, given the state at time $t-1$. We can then follow a similar argument and approximate $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)$ as follows:

$$p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta) \approx \prod_{n\in[1:N]} \sum_{\mathbf{x}_s^{(n)}\in[1:M]} \boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,N\mathbf{m}_{s-1|t}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)}$$
$$=: \prod_{n\in[1:N]} \boldsymbol{\xi}_{n,s-1}^{(\mathbf{x}_{s-1}^{(n)})},$$

where $\boldsymbol{\xi}_{n,s-1}$ is the approximate probability for each individual n of observing the future observation $\mathbf{y}_{[s:t]}$, given the state at time $s-1$. Note that the marginalization is repeated for all states of $\mathbf{x}_{s-1}^{(n)}$, and not $\mathbf{x}_{s-1}$, which reduces the cost from $\mathcal{O}(NM^{N+1})$ to $\mathcal{O}(NM^2)$. We can now build our proposal distribution for an SMC, and approximate $p(\mathbf{x}_s|\mathbf{x}_{s-1},\mathbf{y}_{[s:t]},\theta)$ as

$$p(\mathbf{x}_s|\mathbf{x}_{s-1},\mathbf{y}_{[s:t]},\theta) \approx \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)}}{\tilde{\boldsymbol{\xi}}_{n,s}^{(\mathbf{x}_{s-1}^{(n)})}},$$

$$p(\mathbf{x}_0|\mathbf{y}_{[1:t]},\theta) \approx \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})}}{\tilde{\xi}_{n,0}},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

for $s \in [1:t]$, and with

$$\tilde{\boldsymbol{\xi}}_{n,s}^{(\mathbf{x}_{s-1}^{(n)})} := \sum_{\mathbf{x}_s^{(n)} \in [1:M]} \boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1 - \mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)},$$

$$\tilde{\xi}_{n,0} := \sum_{\mathbf{x}_0^{(n)} \in [1:M]} \boldsymbol{\xi}_{n,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})},$$

for $n \in [1:N]$.

It is crucial to understand the difference between $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$. We use $\boldsymbol{\xi}_{n,s}$ to approximate $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)$ without knowing $\mathbf{x}_s$, and so it is computed by substituting $\mathbf{c}_s$ with $N\mathbf{m}_{s|t}$. We use $\tilde{\boldsymbol{\xi}}_{n,s}$ to approximate $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)$ when we know $\mathbf{x}_s$, and so have access to the actual $\mathbf{c}_s$. The latter is important, because when considering the proposal distribution of an SMC, we know the latest particles, and want to propose the next time step given the last. Note that there are no latest particles in the special case $s = 0$, and hence the recursion looks different; in particular, $\tilde{\xi}_{n,0}$ is a scalar, and can be used to approximate the marginal likelihood $p(\mathbf{y}_{[1:t]}|\theta)$. This approximation could be useful, for pseudo-likelihood methods (Andrieu and Roberts, 2009) or when implemented in a delayed acceptance particle MCMC (Golightly et al., 2015).

We need only $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$ to compute our approximate proposal distribution, and both can be computed before running the SMC at a computational cost $\mathcal{O}(tNM^2)$. However, this also requires a memory cost $\mathcal{O}(tNM^2)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Algorithm 1 Computation of $(\boldsymbol{\xi}_{n,h,s}, \tilde{\boldsymbol{\xi}}_{n,h,s})_{n\in[1:N]}$

Require: $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$, $(\mathbf{q}_{\tilde{s}})_{\tilde{s}\in[s+1:s+h]}$, $(\mathbf{m}_{\tilde{s}|t})_{\tilde{s}\in[s:s+h-1]}$, $\mathbf{y}_{[s+1:s+h]}$,

if $s \neq 0$ **add** $\mathbf{y}_s, \mathbf{c}_{s-1}$

for $n = 1, \ldots, N$ **do**

  $\boldsymbol{\xi}_{n,h,s+h} \leftarrow \mathbf{1}_M$

  for $\tilde{s} = s+h-1, \ldots, s$ **do**

    $\boldsymbol{\xi}^{\mathrm{T}}_{n,h,\tilde{s}} \leftarrow \mathbf{K}^{(\bullet,i)}_{n,N\mathbf{m}_{\tilde{s}|t}} \mathbf{q}^{(i)}_{\tilde{s}+1} \boldsymbol{\xi}^{(i)}_{n,h,\tilde{s}+1} \mathbb{I}_{\mathbf{y}^{(n)}_{\tilde{s}+1}}(i) + \mathbf{K}_{n,N\mathbf{m}_{\tilde{s}|t}} (\mathbf{1}_M - \mathbf{q}_{\tilde{s}+1} \circ \boldsymbol{\xi}_{n,h,\tilde{s}+1}) \mathbb{I}_{\mathbf{y}^{(n)}_{\tilde{s}+1}}(0)$

  if $s \neq 0$ **then**

    $\tilde{\boldsymbol{\xi}}^{\mathrm{T}}_{n,h,s} \leftarrow \mathbf{K}^{(\bullet,i)}_{n,\mathbf{c}_{s-1}} \mathbf{q}^{(i)}_{s} \boldsymbol{\xi}^{(i)}_{n,h,s} \mathbb{I}_{\mathbf{y}^{(n)}_{s}}(i) + \mathbf{K}_{n,\mathbf{c}_{s-1}} \left(\mathbf{1}_M - \mathbf{q}_s \circ \boldsymbol{\xi}_{n,h,s}\right) \mathbb{I}_{\mathbf{y}^{(n)}_{s}}(0)$

  else

    $\tilde{\xi}_{n,h,0} \leftarrow \mathbf{p}^{\mathrm{T}}_{0,n} \boldsymbol{\xi}_{n,h,1}$

because they have to be accessible when running the SMC, which is problematic when t is large. As an alternative, we can compute $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$ at each step of the SMC, which requires a computational cost $\mathcal{O}(Pt^2NM^2)$. A computational cost that is quadratic in t is still undesirable; hence, we can reduce it by using the observations from the closest future instead of those from the whole sequence. We can indeed focus on approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:s+h]}, \theta)$, for $h \in \mathbb{N}$ and $h \ll t$. Given that we have presented our approximation for an arbitrary t, approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:s+h]}, \theta)$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

is like approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$, for $t = s + h$. However, we make the dependence on h explicit by defining $\boldsymbol{\xi}_{n,h,s}$ and $\tilde{\boldsymbol{\xi}}_{n,h,s}$ as $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$, respectively, obtained from the algorithm when looking h steps ahead. The whole procedure is summarized in Algorithm 1, and requires a computational cost $\mathcal{O}(hNM^2)$. Embedding this algorithm in an SMC demands a computational cost $\mathcal{O}(PthNM^2)$, which can be controlled, depending on the computational resources and application. We conclude the section by stating our optimal proposal distribution:

$$q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}, \theta) = \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,h,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1 - \mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)}}{\tilde{\boldsymbol{\xi}}_{n,h,s}^{(\mathbf{x}_{s-1}^{(n)})}},$$

$$q(\mathbf{x}_0|\mathbf{y}_{[1:t]}, \theta) = \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,h,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})}}{\tilde{\xi}_{n,h,0}}.$$

4.3 Resampling

The resampling scheme $(r_s(i))_{s\in[0:t]}$ is not trivial. Indeed, choosing resampling schemes that do not consider future observations negates the effort of building optimal proposals (Fearnhead, 2008). Ideally, resampling should be done according to the smoothing distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:t]}, \theta)$ (Scharth and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Kohn, 2016):

$$p(\mathbf{x}_s|\mathbf{y}_{[1:t]},\theta) = \frac{p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s,\theta)p(\mathbf{x}_s|\mathbf{y}_{[1:s]},\theta)}{p(\mathbf{y}_{[s+1:t]}|\mathbf{y}_{[1:s]},\theta)} \propto p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s,\theta)p(\mathbf{x}_s|\mathbf{y}_{[1:s]},\theta),$$

which is a combination of the probability of observing the future observations, given the current sample $\mathbf{x}_s$, and the filtering distribution. In the equivalent of the low-cost case, the proposal distribution approximation $p(\mathbf{x}_s|\mathbf{x}_{s-1},\mathbf{y}_{[s:s+h]},\theta)$ follows trivially for $t = s+h$.

The quantities involved in the optimal resampling cannot be computed in closed form, and thus need to be approximated. An SMC outputs a particle approximation $(N)^{-1}\sum_{p\in[1:P]} w_s^p \delta_{\mathbf{x}_s^p}(\mathbf{x}_s)$ of the filtering distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:s]},\theta)$, and algorithm 1 yields an approximation $\prod_{n\in[1:N]} \tilde{\boldsymbol{\xi}}_{n,h,s+1}^{(\mathbf{x}_s^{(n)})}$, for $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s,\theta)$. It then follows that the approximate optimal resampling is

$$r_s(i) \propto w_s^i \prod_{n\in[1:N]} \tilde{\boldsymbol{\xi}}_{n,h,s+1}^{((\mathbf{x}_s^i)^{(n)})} \quad \text{for } i \in [1:P],$$

with $(\mathbf{x}_s^p)_{p\in[1:P]}$ being the sampled particles at time s.

5. Experiments

In this section, we analyze the performance of SMC algorithms when using our approximation of the optimal proposal and resampling scheme. We consider simulated data from the SIS model and the SEIR model, which

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

we analyze in Section 5.1 and Section 5.2, respectively. For each model, we follow an experimental routine inspired by Ju et al. (2021), to compare our method with the BPF and APF: (i) we compare the methods based on the effective sample size (ESS) $1/\sum_{i\in[1:P]}(r_s(i))$; (ii) we compare the methods based on the standard deviation of the estimate of marginal likelihood; and (iii) we study the marginal likelihood surface on a grid of parameter values for different t when using our method.

All experiments are run on a 32 GB Tesla V100 GPU, made available from the HEC (High-End Computing) facility from Lancaster University. The code is available in the GitHub repository "Optimal_IBM_proposal" (https://github.com/LorenzoRimella/Optimal_IBM_proposal).

5.1 The SIS model

The SIS model is used in epidemiology to model the spread of a disease in a population when herd immunity is not possible. As already mentioned in Section 3.1, we can formulate an individual-based model by having individuals specific covariates $(\mathbf{w}_n)_{n\in[1:N]}$, and using these covariates to define a unique dynamic per individual. We have covariates of the form $\mathbf{w}_n = [\mathbf{w}_n^{(1)}, \mathbf{w}_n^{(2)}]^{\mathrm{T}}$, where $\mathbf{w}_n^{(1)} = 1$ and $\mathbf{w}_n^{(2)} \sim \mathcal{N}(\bullet|0,1)$ independently, for all $n \in [1:N]$. If not specified otherwise, we consider $N = 100$, the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

time horizon $t = 100$, and data-generating parameters (DGP) given by $\boldsymbol{\beta_0} = [-\log(N-1), 0]^{\mathrm{T}}$, $\boldsymbol{\beta_\lambda} = [-1, 2]^{\mathrm{T}}$, $\boldsymbol{\beta_\gamma} = [-1, -1]^{\mathrm{T}}$, and $\mathbf{q}_t = \mathbf{q}$, with $\mathbf{q} = [0.8, 0.8]^{\mathrm{T}}$.

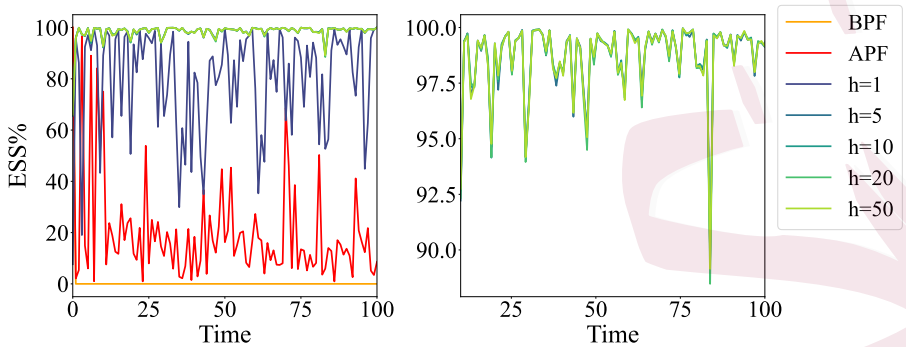

Figure 2: ESS percentage over time for BPF, APF, and our method when $h = 1, 5, 10, 20, 50$. The left plot shows all the listed methods, and the right plot considers only $h = 10, 20, 50$ and zoom-in.

The first experiment measures the ESS for the BPF, the APF, and $h = 1, 5, 10, 20, 50$ when $P = 512$. Figure 2 displays our findings for a number of particles $P = 512$. The BPF fails in terms of sampling any epidemics trajectories, because of the mismatch problem mentioned in Section 3.3, where a single mismatch out of N individuals is sufficient to assign a zero probability to the associated particle. The APF corrects the proposal by considering the current observation, and so avoids a mismatch. Even though this is a significant improvement on the BPF, the ESS is still very low.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Our approximate optimal proposal reaches a significantly better ESS than that of the APF by just looking at the next step in the future ($h = 1$). Furthermore, choosing $h > 5$ does not improve the performance, owing to the forgetting property of our HMM (Douc et al., 2009).

In the next experiment, we examine the standard deviation of the marginal likelihood estimates. We consider two frameworks: one using the data-generating parameters, and the other substituting β_λ with $[-3, 0]^{\mathrm{T}}$. Standard deviations are computed over 100 runs. The APF is three to four times faster than our method when $h = 5$, but the standard deviation in both frameworks is 10 to 20 times higher than $h = 5$ for small P, and 20 to 30 times higher than $h = 5$ for big P. Again, we find no substantial improvement when using $h > 5$. The computational cost depends highly on the implementation. Our scripts run on GPUs and parallelize each step of the SMC across individuals and particles. Hence we do not report significant changes in the running time when increasing P.

Now, suppose we want to infer β_λ or β_γ. We start by setting β_λ in a two-dimensional grid on $[-4, 4]^2$, and the other parameters to the data-generating parameters (including β_γ). We then compute estimates of the marginal likelihood with an SMC using our proposal and a resampling scheme when $P = 512$. The procedure is replicated for β_γ. Both experi-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Table 2: The standard deviation for the APF and our method when $h = 5, 10, 20$ under the data-generating process (DGP) and non-data-generating process (NDGP) with $P = 128, 512, 2048$. The mean computational cost of a single step of the SMC is reported in the first row with the name of the algorithm.

	APF	0.7s	h=5	2.5s	h=10	3.94s	h=20	6.61s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
128	4.99	9.89	0.3	0.92	0.31	1.0	0.37	0.89
512	4.01	6.66	0.17	0.48	0.18	0.49	0.18	0.48
2048	2.83	6.23	0.11	0.25	0.11	0.22	0.11	0.22

ments are run for $h = 5, 10$ and $t = 50, 100$, with new data generated for each value of t. Marginal likelihood contour plots are reported in Figure 3 in log scale, and are normalized to have their maximum in zero.

In both figures, increasing the time concentrates the likelihood around the data-generating parameters. Choosing $h = 10$ does not improve the inference over β_λ, but it helps to infer β_γ by removing some combinations of parameters from the inference (white space).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

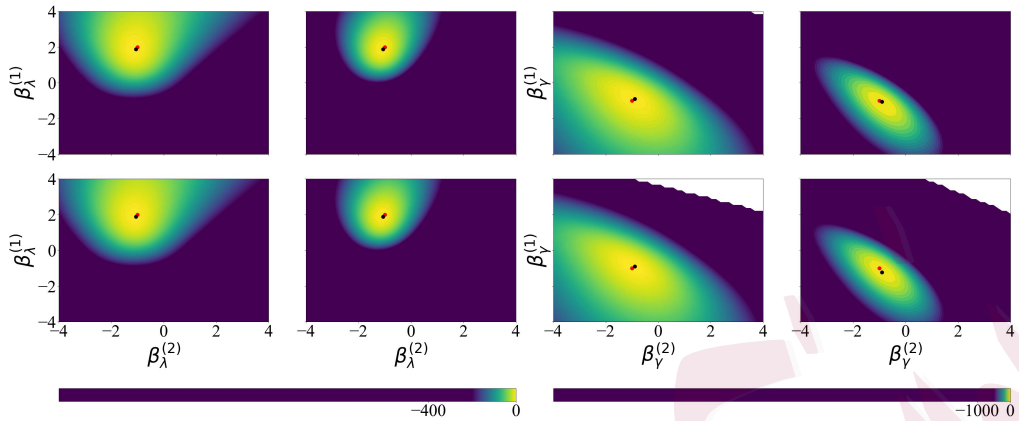

Figure 3: Marginal likelihood contour plots on a β_λ grid and a β_γ grid in log-scale. The first and second columns refer to $t = 50, 100$, from left to right, for β_λ. The third and fourth columns refer to $t = 50, 100$, from left to right, for β_γ. Rows refer to $h = 5, 10$, from top to bottom. The colorbars are common across parameters, and their maximum is set to zero. In red, are the data-generating parameters, and in black are the MLEs on the grid.

5.2 The SEIR model

The SEIR model is popular in epidemiology (He et al., 2020; Deguen et al., 2000; Porter and Oleson, 2013) when the disease is expected to have a latent period (exposed compartment) and herd immunity (removed compartment). The SEIR case is significantly more challenging than the SIS, because the transition kernel constrains the dynamic on $S \to E \to I \to R$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

and so if in our SMC at time $t-1$ we have a particle with individual n in compartment S, and we then observe the same individual at time t in compartment I or R, the SMC assigns a zero probability to that particle.

As for the SIS case, a heterogeneous SEIR model is obtained by including a collection of covariates defining $(\mathbf{w}_n)_{n\in[1:N]}$. The initial distribution $\mathbf{p}_{n,0}$ is defined on compartments 1 (S) and 3 (I), as for the SIS case, with zeros for compartments 2 (E) and 4 (I). Similarly, $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ is defined as the SIS for transitions $1,2$ (S, E) and $3,4$ (I, R), with the additional transition $2,3$ (E, I) given by $1-\exp(-\rho)$. Full definitions of $\mathbf{p}_{n,0}$ and $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ are available in the Supplementary Material. The emission distribution follows (3.1).

We have covariates of the form $\mathbf{w}_n = [\mathbf{w}_n^{(1)}, \mathbf{w}_n^{(2)}]^{\mathrm{T}}$, where $\mathbf{w}_n^{(1)} = 1$ and $\mathbf{w}_n^{(2)} \sim \mathcal{N}(\bullet|0,1)$ independently, for all $n \in [1:N]$. If not specified otherwise, we consider $N = 1000$, the time horizon $t = 100$, and data-generating parameters given by $\boldsymbol{\beta_0} = [-\log(N/10-1), 0]^{\mathrm{T}}$, $\boldsymbol{\beta_\lambda} = [1,2]^{\mathrm{T}}$, $\rho = 0.2$, $\boldsymbol{\beta_\gamma} = [-1,-1]^{\mathrm{T}}$, and $\mathbf{q}_s = \mathbf{q}$, with $\mathbf{q} = [0,0,0.4,0.6]^{\mathrm{T}}$.

As for the SIS case, we start by analyzing the ESS for the BPF, the APF, and $h = 1,5,10,20,50$, with $P = 512$. In Figure 4, both the BPF and the APF fail because of a mismatch between the proposed particles and the observations. Even our method fails for $h \leq 5$, but when choosing $h \geq 10$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

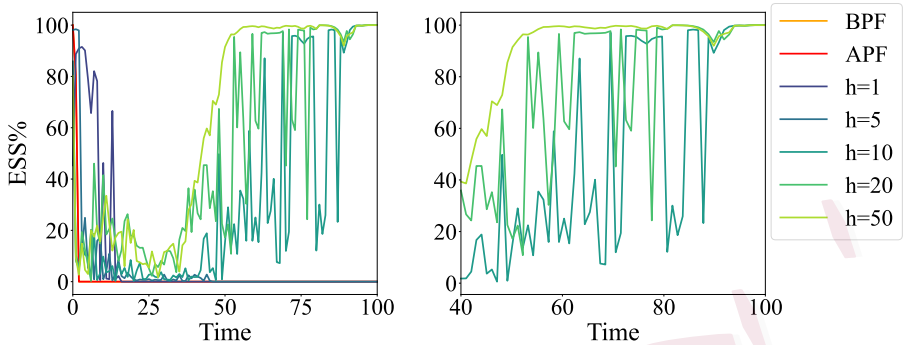

Figure 4: ESS percentage over time for the BPF, the APF, and our method when $h = 1, 5, 10, 20, 50$. The left plot shows the listed methods, and the right plot considers only $h = 5, 10, 20, 50$ and zoom-in.

we avoid a mismatch, and get an increasing in time ESS.

We then investigate the standard deviation and computational cost of our method when h changes, and report our results in Table 3. There is a significant improvement in the standard deviation when increasing h to 50, with the jump from $h = 20$ to $h = 50$ being less substantial. Clearly, there is a trade-off, because a decrease in the standard deviation increases the computational cost, which appears to be worth it for $h < 50$, because halving the standard deviation less than doubles the computational cost.

We conclude by reproducing the marginal likelihood surfaces of β_λ and β_γ on grids for the SEIR scenario. The experiments are run for $h = 10, 20$ and $t = 50, 100$, with new data generated for each value of t. Figure 5 shows

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Table 3: The standard deviation for our method when $h = 5, 10, 20, 50$ under the data-generating parameters (DGP) and non-data-generating parameters (NDGP) with $P = 128, 512, 2048$. The mean computational cost of a single step of the SMC is reported in the first row with the name of the algorithm.

	h=5	0.9s	h=10	3.5s	h=20	5.45s	h=50	9.03s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
128	58.18	68.2	20.47	32.6	9.59	18.32	6.93	11.71
512	48.23	74.78	18.37	28.64	6.39	15.76	6.23	10.72
2048	42.7	58.37	15.03	24.68	5.69	13.25	4.57	10.45

the marginal likelihood contour plots on a log scale, and are normalized to have their maximum in zero. As in the SIS case, an increase in t concentrates the likelihood around the DGP, as shown in both figures. Note that the log-likelihood surface of β_λ is multi-modal, because we observe individuals who are neither susceptible nor exposed, which makes inference on this parameter significantly harder. Choosing $h = 20$ seems to smooth the likelihood surface, and avoids failure close to the data-generating parameters, as seen by the white holes in the surface for $h = 10$ and $t = 100$. In addition, β_γ has a smoother surface than that of β_λ, and increasing h seems

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

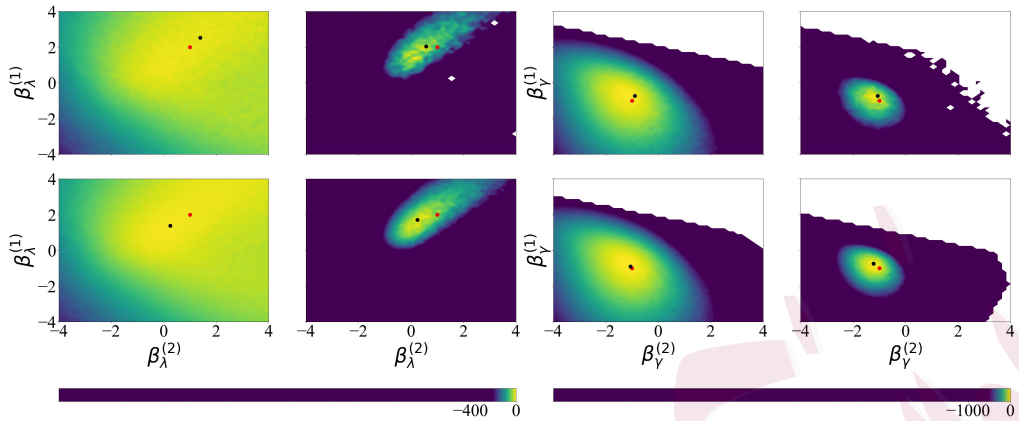

Figure 5: Marginal likelihood contour plots on a β_λ grid and a β_γ grid in log-scale. The first and second columns refer to $t = 50, 100$ from left to right for β_λ. The third and fourth columns refer to $t = 50, 100$ from left to right for β_γ. Rows refer to $h = 5, 10$ from top to bottom. The colorbars are common across parameters and their maximum is set to 0. In red are the data-generating parameters and in black are the MLE on the grid.

to improve the shape.

6. Discussion

Our findings demonstrate the difficulties in fitting individual-based epidemic models in the presence of censored data, and highlight the significance of incorporating future observations when choosing proposal distributions

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

in SMC algorithms. The underlying framework in which proposal distributions is general, and the algorithm requires only that we obtain, estimates of the transition rates at the times $[t+1, t+h]$ for each individual, which are then propagated backward to build a proposal distribution that includes future observations.

Although the proposed procedure computes good proposal distributions, several aspects can be used to improve existing methods. For example, our backward recursion method can be used to compute the marginal likelihood approximation in pseudo-likelihood methods (Andrieu and Roberts, 2009), or as the first approximate model stage in the delayed acceptance scheme of Golightly et al. (2015).

Our implementation focuses on the case of homogeneous reporting rates in a fully connected population, but can be extended to heterogeneous reporting rates, as discussed at the end of Section 4.1. For spatial epidemic models, we simply need to obtain an estimate of the spatial risk of infection to be able to run the recursion. Epidemic models with an open population (e.g., migration or births-deaths) and misreporting can also be included in the class of models we deal with by substituting the multinomial approximation (Whiteley and Rimella, 2021) with alternative approximations (Whitehouse et al., 2022).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Supplementary Material

The online supplementary material is divided into five sections:

1. the main notation and conventions;
2. an introduction to the compartmental model;
3. the main computation in the HMM and SMC;
4. the algorithm of Whiteley and Rimella (2021); and
5. additional experiments and extra details on some of the experiments presented in the main paper.

Acknowledgments

This work was supported by EPSRC grants EP/R018561/1 (Bayes4Health) and EP/R034710/1 (CoSInES). The authors thank Simon Spencer for a useful discussion, that prompted the idea for this research.

References

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Bickel, P., B. Li, and T. Bengtsson (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pp. 318–329. Institute of Mathematical Statistics.

Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences 225*(1), 24–35.

Brooks-Pollock, E., L. Danon, T. Jombart, and L. Pellis (2021). Modelling that shaped the early COVID-19 pandemic response in the UK. *Phil. Trans. R. Soc. B 376*, 20210001.

Candy, J. V. (2007). Bootstrap particle filtering. *IEEE Signal Processing Magazine 24*(4), 73–85.

Carpenter, J., P. Clifford, and P. Fearnhead (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation 146*(1), 2–7.

Chapman, L. A. C., S. E. F. Spencer, T. M. Pollington, C. P. Jewell, D. Mondal, J. Alvar, T. D. Hollingsworth, M. M. Cameron, C. Bern, and G. F. Medley (2020). Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis. *Proceedings of the National Academy of Sciences 117*(41), 25742–25750.

Chopin, N., O. Papaspiliopoulos, et al. (2020). *An introduction to sequential Monte Carlo*, Volume 4. Springer.

Cocker, D., M. Sammarro, K. Chidziwisano, N. Elviss, S. Jacob, H. Kajumbula, L. Mugisha, D. Musoke, P. Musicha, A. Roberts, B. Rowlingson, A. Singer, R. Byrne, T. Edwards,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

R. Lester, C. Wilson, B. Hollihead, N. Thomson, C. Jewell, T. Morse, and N. Feasey (2022, February). Drivers of Resistance in Uganda and Malawi (DRUM): a protocol for the evaluation of One-Health drivers of Extended Spectrum Beta Lactamase (ESBL) resistance in Low-Middle Income Countries (LMICs). *Wellcome Open Res.*.

Deguen, S., G. Thomas, and N. P. Chau (2000). Estimation of the contact rate in a seasonal SEIR model: application to chickenpox incidence in France. *Statistics in Medicine 19*(9), 1207–1216.

Douc, R., E. Moulines, and Y. Ritov (2009). Forgetting of the initial condition for the filter in general state-space hidden Markov chain: a coupling approach. *Electronic Journal of Probability 14*, 27–49.

Fearnhead, P. (2008). Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Statistics and Computing 18*(2), 151–171.

Funk, S., S. Abbott, B. Atkins, M. Baguelin, J. Baillie, P. Birrell, J. Blake, N. Bosse, J. Burton, J. Carruthers, N. Davies, D. De Angelis, L. Dyson, W. Edmunds, R. Eggo, N. Ferguson, K. Gaythorpe, E. Gorsich, G. Guyver-Fletcher, J. Hellewell, E. Hill, A. Holmes, T. House, C. Jewell, M. Jit, T. Jombart, I. Joshi, M. Keeling, E. Kendall, E. Knock, A. Kucharski, K. Lythgoe, S. Meakin, J. Munday, P. Openshaw, C. Overton, F. Pagani, J. Pearson, P. Perez-Guzman, L. Pellis, F. Scarabel, M. Semple, K. Sherratt, M. Tang, M. Tildesley, E. Van Leeuwen, L. Whittles, C. C.-. W. Group, I. C. C.-. R. Team, and I. Investigators (2020). Short-term forecasts to inform the response to the COVID-19 epidemic in the UK.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

medRxiv.

Golightly, A., D. A. Henderson, and C. Sherlock (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing 25*(5), 1039–1055.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-Radar and Signal Processing*, Volume 140, pp. 107–113. IET.

He, S., Y. Peng, and K. Sun (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics 101*(3), 1667–1680.

Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

Jewell, C. P., M. J. Keeling, and G. O. Roberts (2009, Dec). Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *J R Soc Interface 6*(41), 1145–1151.

Jewell, C. P., T. Kypraios, P. Neal, and G. O. Roberts (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis 4*(3), 465–496.

Johansen, A. M. and A. Doucet (2008). A note on auxiliary particle filters. *Statistics & Probability Letters 78*(12), 1498–1504.

Ju, N., J. Heng, and P. E. Jacob (2021). Sequential Monte Carlo algorithms for agent-based models of disease transmission. arXiv:2101.12156.

Kucharski, A. J., T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

F. Sun, M. Jit, J. D. Munday, et al. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases 20*(5), 553–558.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*(446), 590–599.

Porter, A. T. and J. J. Oleson (2013). A path-specific SEIR model for use with general latent and infectious time distributions. *Biometrics 69*(1), 101–108.

Rabiner, L. and B. Juang (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine 3*(1), 4–16.

Rimella, L., M. Sammarro, S. Alderton, N. Feasey, P. Fearnhead, and C. Jewell (2022). Inference on extended-spectrum beta-lactamase *Escherichia coli* and *Klebsiella pneumoniae* data through SMC^2. *In Preparation*.

Scharth, M. and R. Kohn (2016). Particle efficient importance sampling. *Journal of Econometrics 190*(1), 133–147.

Van der Goot, J., G. Koch, M. De Jong, and M. Van Boven (2005). Quantification of the effect of vaccination on transmission of avian influenza (H7N7) in chickens. *Proceedings of the National Academy of Sciences 102*(50), 18141–18146.

Whitehouse, M., N. Whiteley, and L. Rimella (2022). Consistent and fast inference in compartmental models of epidemics using poisson approximate likelihoods. *arXiv preprint arXiv:2205.13602*.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0198

Whiteley, N. and A. Lee (2014). Twisted particle filters.

Whiteley, N. and L. Rimella (2021). Inference in stochastic epidemic models via multinomial approximations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1297–1305. PMLR.

Yang, F., S. Balakrishnan, and M. J. Wainwright (2017). Statistical and computational guarantees for the Baum-Welch algorithm. *The Journal of Machine Learning Research 18*(1), 4528–4580.

Zhou, Y., Z. Ma, and F. Brauer (2004). A discrete epidemic model for SARS transmission and control in China. *Mathematical and Computer Modelling 40*(13), 1491–1506.

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: l.rimella@lancaster.ac.uk

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: c.jewell@lancaster.ac.uk

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: p.fearnhead@lancaster.ac.uk